Automatic or Manual? A Fuel Consuption Analysis

Oszkar Jozsa

Summary

In this paper, I will show that, in general, cars with automatic transmission have worse fuel economy compared to the manual transmission ones. However, my analysis also will show that the worse mileage performance is determined by other factors also, not only the type of transmission the car has.

After a quick analysis in the first section, I will construct a linaer prediction model that predicts the fuel consumption based on those factors.

Analysis

Data processing

First of all, let's load the data and have a quick look at it.

```
library(ggplot2)
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

We have 32 observations. Our two important variables are `mpg` and `am`. First one is a numeric value of the fuel consumption, second one is the gear shift type. This site tells us that 0 = automatic and 1 = manual.

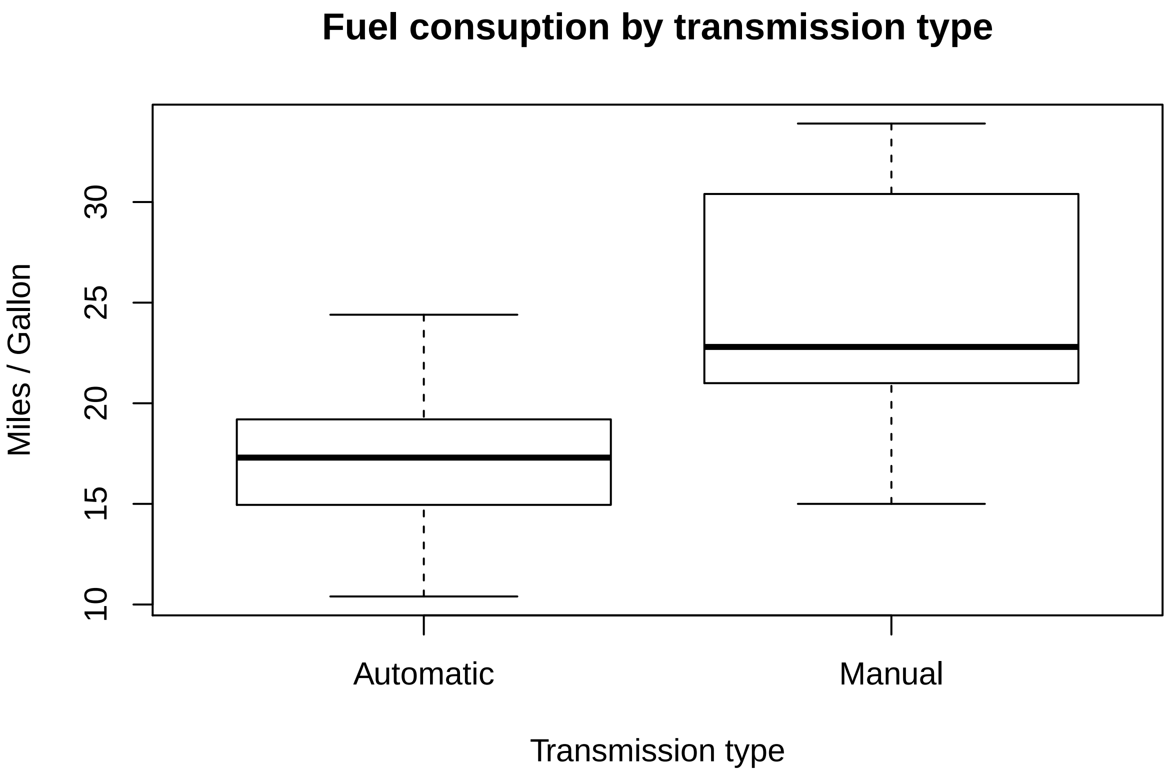
Lat's convert that numerical value into a factor:

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Exploratory analysis

First of all, check if there is any difference between automatic and manual transmission with respect to MPG at all. A boxplot can clearly visualize this for us.

```
boxplot(mpg~am, data=mtcars,
        xlab = "Transmission type",
        ylab = "Miles / Gallon",
        main = "Fuel consuption by transmission type")
```



Vehicles with automatic transmission seem to have lower MPG, or in other words, they have worse fuel consumption.

Another validation of this initial assumption is to break down these categories further. It's a good intuition that engine displacement and weight can effect fuel consuption, we should check whether these parameters are similar in case of automatic and manual transmissions.

```
summary(mtcars[mtcars$am == "Automatic",]$disp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   120.1   196.3   275.8   290.4   360.0   472.0
```

```
summary(mtcars[mtcars$am == "Manual",]$disp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.1    79.0   120.3   143.5   160.0   351.0
```

```
summary(mtcars[mtcars$am == "Automatic",]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.465   3.438   3.520   3.769   3.843   5.424
```

```
summary(mtcars[mtcars$am == "Manual",]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.513   1.935   2.320   2.411   2.780   3.570
```

It shows that automatic shift cars tend to have larger engine displacement and larger weight, too. This suggests that the larger fuel consumption of automatic shift cars might not caused by the shift type itself. Of course, other factors can effect fuel consuption but most likely we have to include these variables also on our final predicition model. Appendix 1 and Appendix 2 show the displacement and weight versus MPG per transmission type diagrams.

One final thing we can check to make sure the data is valid: it doesn't seem to have outliers and it seems to be roughly normally distributed. So this is not skewing our data either. See Appendix 3 for the distribution analysis.

Regression model

Simple regression

The simplest model is to use a single variable regression, using the transmission type, `am`.

```
singlelm <- lm(mpg~am, data=mtcars)
summary(singlelm)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Turns out (unsurprisingly) that the single variable of the transmission type is not that great of a predictor of fuel consumption. We can see that using an automatic transmission, our fuel consumption tends to be 7 MPG higher compared to manual transmission. However, the Rˆ2 value tells us that this model only explains 36% of the variance. Also, the squared error seems to be high, however, p-value is under 0.05.

Multivariare regression

I will use a 2-way stepwise selection to build the model. Details of stepwise regression: [WikiPedia](), [stack exchange](), [Coastal Carolina University tutorial]().

```
stepmodel = step(lm(data=mtcars, mpg~.), trace=0, steps=10000)
summary(stepmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

So it seems that the transmission type is only the third best predictor variable in the final model. Both weight and acceleration are much better predictors (the latter one strongly relates to displacement size and cylinder number).

```
multiplelm <- lm(mpg ~ wt + qsec + am, data=mtcars)
summary(multiplelm)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

In the final model, the R^2 value is up to 85% and the p-value decreased well belo 0.05.

Results

We have found a multiple variable regression model that covers 85% of the variable in the fuel consumption, based on the `mtcars` dataset.

The three variables in the model:

- wt - the weight of the car
- qsec - acceleration time (0.25 mile)
- am - transmission type (automatic/manual)

The performance of this model compared to a single variable model:

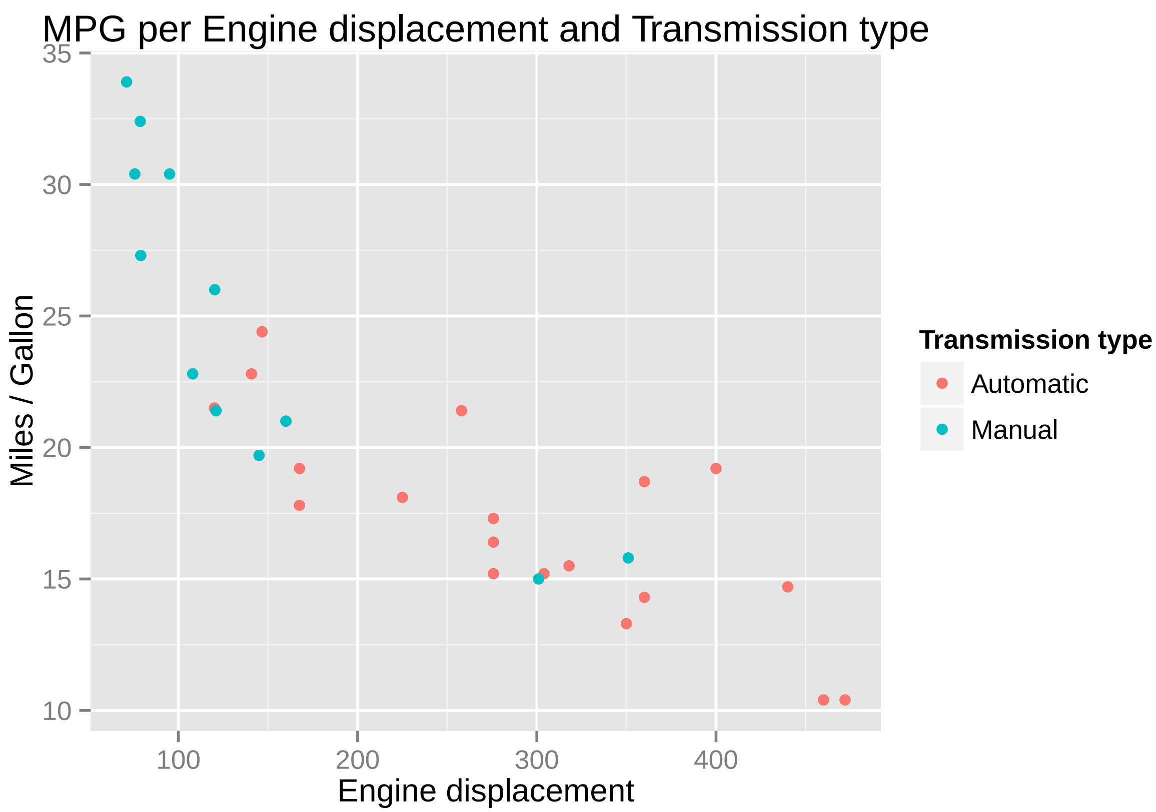
```
rmse_single <- sqrt(sum(singlelm$residuals ^ 2) / nrow(mtcars))
rsquared_single <- summary(singlelm)$r.squared
rmse_multiple <- sqrt(sum(multiplelm$residuals ^ 2) / nrow(mtcars))
rsquared_multiple <- summary(multiplelm)$r.squared
perf <- matrix(c(rmse_single, rmse_multiple, rsquared_single, rsquared_multiple), nrow=2, ncol=2)
rownames(perf) <- c("single variable", "multiple variable")
colnames(perf) <- c("RMSE", "R^2")
perf
```

```
##                       RMSE       R^2
## single variable   4.746369 0.3597989
## multiple variable 2.300040 0.8496636
```

Appendix

Appendix 1 - Engine displacement

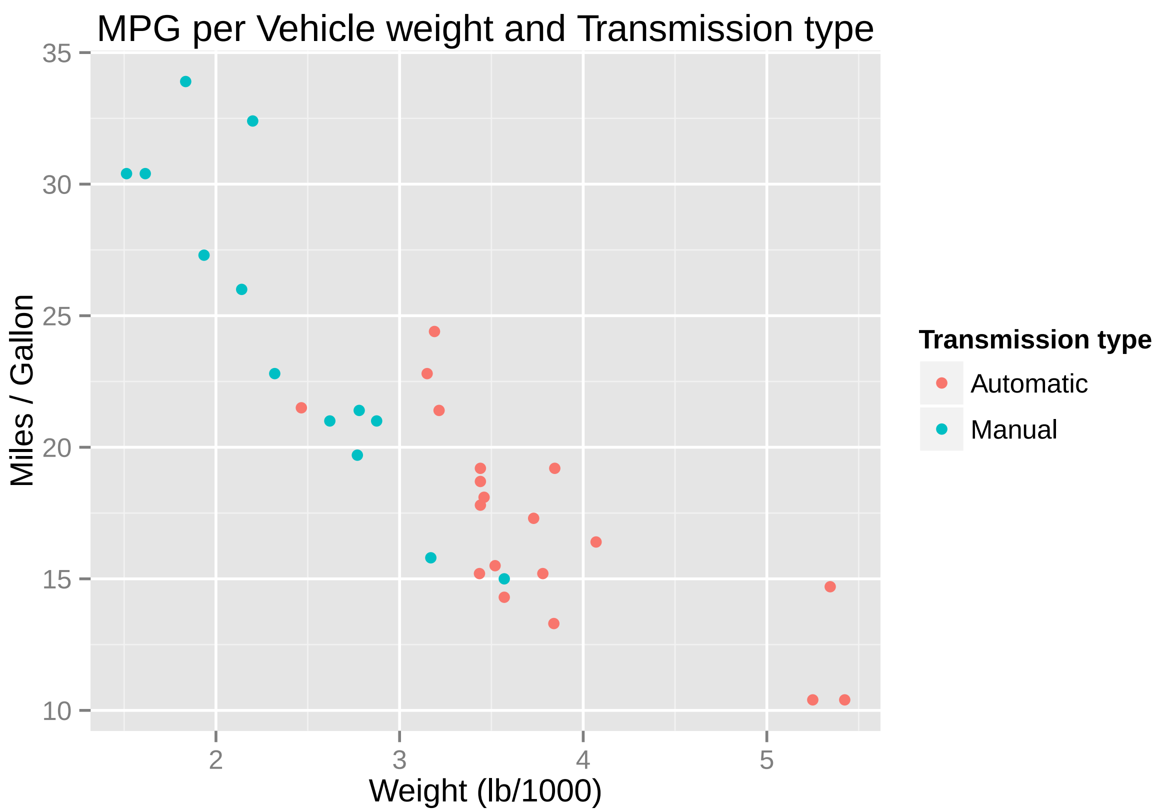
```
dispplot <- qplot(x=disp, y=mpg, data=mtcars, colour=am,
                  main="MPG per Engine displacement and Transmission type",
                  xlab="Engine displacement",
                  ylab="Miles / Gallon")
dispplot + labs(colour="Transmission type")
```



It can be seen that even though automatic shift cars have bigger engines and bigger fuel consuption also.

Appendix 2 - Vehicle weight

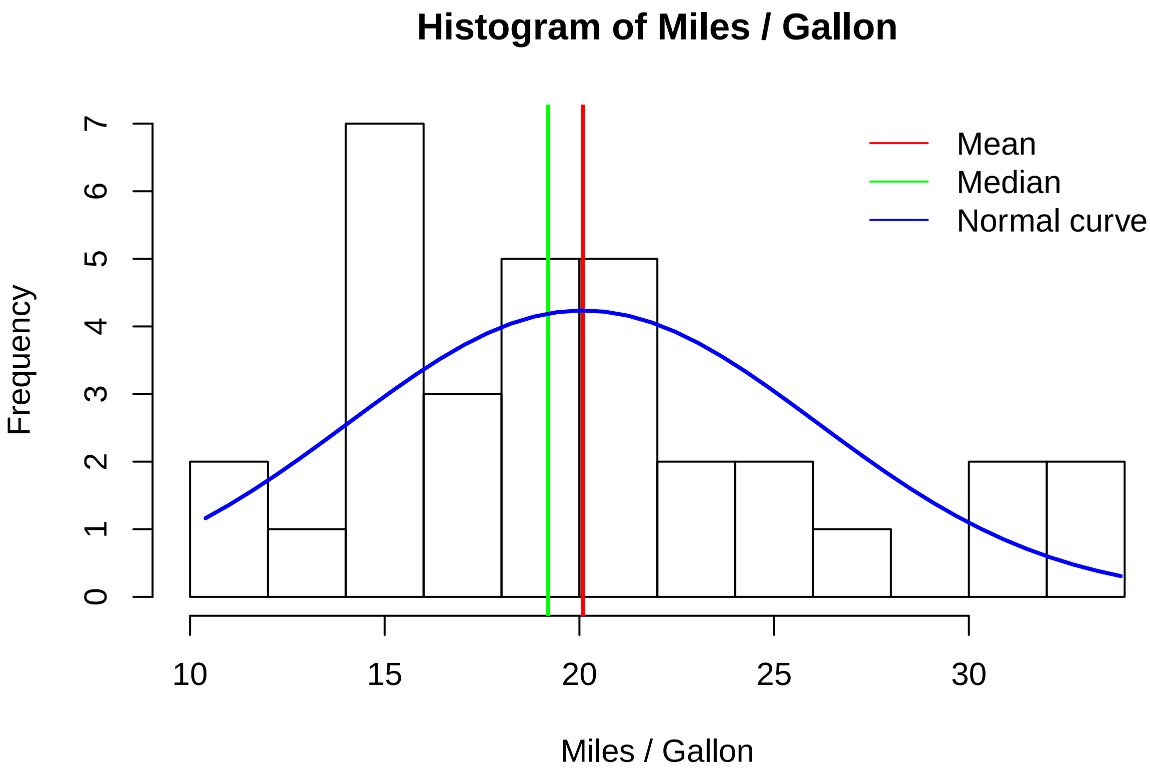
```
wtplot <- qplot(x=wt, y=mpg, data=mtcars, colour=am,
                main="MPG per Vehicle weight and Transmission type",
                xlab="Weight (lb/1000)",
                ylab="Miles / Gallon")
wtplot + labs(colour="Transmission type")
```



It can be seen that even though automatic shift cars are heavier and have bigger fuel consumption.

Appendix 3 - MPG distribution

```
mpg <- mtcars$mpg
h <- hist(mpg, breaks=11, xlab="Miles / Gallon",  main="Histogram of Miles / Gallon")
abline(v = mean(mpg), col="red", lwd=2)
abline(v = median(mpg), col="green", lwd=2)
xfit <- seq(min(mpg), max(mpg), length=40)
yfit <- dnorm(xfit, mean=mean(mpg), sd=sd(mpg))
yfit <- yfit * diff(h$mids[1:2]) * length(mpg)
lines(xfit, yfit, col="blue", lwd=2)
legend(x="topright", legend=c("Mean","Median","Normal curve"), col=c("red", "green","blue"), bty="n", lwd=1)
```


Histogram of Miles / Gallon
Mean
Median
Normal curve
Frequency
0
1
2
3
4
5
6
7
10
15
20
25
30
Miles / Gallon